DOCUMENT RESUME

ED 411 296 TM 027 310

AUTHOR	Ruthven, Kenneth
TITLE	Beyond Common Sense: Reconceptualizing National Curriculum Assessment.
ISSN	ISSN-0958-5176
PUB DATE	1995-00-00
NOTE	28p.
AVAILABLE FROM	Routledge Subscriptions Department, North Way, Andover, Hants SP10 5BE, United Kingdom (US institutions, annual subscription, $100; US individuals, annual subscription, $60; single issues and back copies, 20 pounds, $35 US).
PUB TYPE	Journal Articles (080) -- Reports - Evaluative (142)
JOURNAL CIT	Curriculum Journal; v6 n1 p5-28 Spr 1995
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*British National Curriculum; Curriculum Development; *Educational Assessment; Educational Research; Elementary Secondary Education; Foreign Countries; *Learning; *Mathematics; Scaling; Student Evaluation; Tables (Data); *Test Construction
IDENTIFIERS	*England; Wales

ABSTRACT

This article examines the implications of the shift, in England and Wales, from the previous National Curriculum assessment model based on the assessment of performance in relation to statements of attainment, to a model based on level descriptions founded on the recommendations of R. Dearing (1994). Both models reflect a common sense view that learning is marked by the stepwise acquisition of curricular objectives, but two key shifts in conceptualization have taken place. One is the move from a focus on actual performance on individual objectives to idealized performance on complexes of objectives, and the other is from a concept of hierarchical progression to one of cumulative progression. This article highlights some of the central weaknesses of the common sense view, in both its original and reviewed versions, by reference to the findings of research on mathematics. Neither model takes sufficient account of the extent to which development and learning in mathematics involve progressive, but uneven, consolidation, coordination, and reorganization of knowledge, with stability of performance achieved only at the end of this process. A more fruitful, but controversial, way to move forward would be to conceptualize a pupil's learning as taking place in a development band, in which concepts are still being refined. The creation of such a system would demand the improved definition of attainment targets and scaling within them and the development of complementary regular and standard assessments. Reliance on politically expedient short tests is unlikely to prove either trustworthy or informative. (Contains 10 tables and 47 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

THE CURRICULUM JOURNAL Vol. 6 No. 1

ED 411 296

Beyond common sense: reconceptualizing National Curriculum assessment

KENNETH RUTHVEN
Department of Education, University of Cambridge

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Kenneth Ruthven

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM 027 310

The Curriculum Journal Vol 6 No 1 Spring 1995 5–28
© The Curriculum Association 1995
ISSN 0958-5176

BEST COPY AVAILABLE

THE CURRICULUM JOURNAL Vol. 6 No. 1

Beyond common sense: reconceptualizing National Curriculum assessment

KENNETH RUTHVEN
Department of Education, University of Cambridge

ABSTRACT

This article examines the implications of the shift, in England and Wales, from the pre-Dearing National Curriculum assessment model based on the assessment of performance in relation to statements of attainment, to the post-Dearing model based on level descriptions. Both the original and revised models reflect a common sense view that learning is marked by stepwise acquisition of curricular objectives, with assessment providing robust and interpretable indices of such progress. However, two key shifts in conceptualization have taken place: from a focus on actual performance on individual objectives to idealized performance on complexes of objectives; and from a concept of hierarchical progression to one of cumulative progression. The article highlights some of the central weaknesses of the common sense view, in both its original and revised versions, by reference to the findings of research on assessment in mathematics. Neither model takes sufficient account of the extent to which development and learning in mathematics involve progressive, but uneven, consolidation, co-ordination and reorganization of knowledge, with stability of performance achieved only at the end of this process. A more fruitful, but controversial, way forward would be to conceptualize a pupil's learning as taking place in a *development band*, in which concepts are still being refined. This would recognize that performance is less consistent in this band, and that task content and presentation have an important influence. The creation of such a system would demand the improved definition of attainment targets, and scaling within them, through further research; and the development of teacher assessment and standard assessment so that they can be genuinely complementary. Reliance on politically expedient short tests is unlikely to prove either trustworthy or informative.

KEY WORDS

assessment; National Curriculum; mathematics education; progression.

The Curriculum Journal Vol 6 No 1 Spring 1995 5–28
© The Curriculum Association 1995
ISSN 0958-5176

TM 027 310

THE COMMON SENSE VIEW OF ASSESSMENT

'It's really all common sense' (Dearing, 1994b) was the modest disclaimer of the author of the skilfully pragmatic recommendations for the revision of the National Curriculum and its assessment (Dearing, 1993, 1994a). More contentiously, in accepting these recommendations, the government (Department for Education, 1994) linked them to its 'back to basics' policies aimed at countering 'fashionable theories [which] are a denial of common sense' (Major, 1994). The appeal to common sense is, however, double-edged. On the one hand, a system of public assessment should certainly be readily comprehensible in its principles, and realistically implementable in its practice; on the other, the uncritical acceptance of popular preconceptions, and the pragmatic search for operational simplicity, may produce a system whose superficial merits mask its inadequate conceptualization and unfortunate effects.

When he commissioned the original arrangements for the National Curriculum and its assessment in 1987, the then Education Secretary sought a system in which 'Attainment targets will provide objectives against which pupils' progress and performance can be assessed . . . to show what a pupil has learnt and mastered, so as to enable teachers and parents to ensure that he or she is making adequate progress and to inform decisions about the next steps' (Task Group on Assessment and Testing, 1988a: Appendix 1). Seeking a thoroughgoing review of these arrangements, six years and three Education Secretaries later, his distant successor maintained this position: 'We must ensure that children's progress is measured clearly in relation to the targets set out in the curriculum so that the next steps for children's learning are apparent to teachers and parents' (Dearing, 1993: Annex 1). Such pronouncements indicate the resilience of the common sense view that learning is marked by stepwise acquisition of curricular objectives, with assessment providing robust and interpretable indices of such progress.

There are operational as well as rhetorical continuities between the old National Curriculum assessment model and the new. Most important is the retention of the structural framework for organizing curricular objectives originating in the recommendations of the Task Group on Assessment and Testing (TGAT, 1988a, 1988b). Within this framework, cognate objectives are grouped to form a limited number of *attainment targets*, and then scaled at ten *levels of performance* according to expected patterns of educational development. Beneath these superficial continuities, however, there are important differences between the assessment models. The central issues are those of progression and measurement. For Dearing: 'In theory, the [ten-level] scale offers an obvious and straightforward statement of progression. A pupil begins at level one and moves, in so far as his/her ability allows, to level 10. In practice, there are serious problems with the attempt to define a pupil's increasing mastery of knowledge, understanding and skills in terms of the statements of attainment which characterize the different levels' (Dearing, 1994a: 60–1).

Effectively, while preserving both the common sense view of assessment and the established curricular framework, Dearing has reformulated the relationship between their elements. Two key shifts in conceptualization have taken place. The first is from *denotative* to *connotative* measurement, signalled by the refocusing of attention from actual performance on individual objectives (referred to as *statements of attainment*) to idealized performance on the complexes of objectives associated with particular levels of an attainment target (referred to as *level descriptions*). In the original model, assessment results were intended to 'give direct information about pupils' achievement in relation to objectives' (TGAT, 1988a: para. 5). For Dearing: 'Achievement at any level needs ... to be judged in the round rather than by reference to a closely specified definition expressed through detailed statements of attainment' (Dearing, 1993: 26). Consequently, in the revised model, the level awarded is that which 'best fits the pupil's performance', and level descriptions 'describe the types and range of performance which pupils working at a particular level should characteristically demonstrate' (SCAA, 1994a: 26).

The second and associated shift is from *hierarchical* to *cumulative* progression. In the original model: 'the sequence of levels represents the stages of progression. A pupil assessed as having achieved a given level, say level 2, will have satisfied the criteria for level 2 and will be working towards the criteria for level 3. Progress is marked by achievement of successive levels over time' (TGAT, 1988a: paras 100–1). In the revised model, while progression is still signified by increasing level, it is no longer tied to the achievement of particular objectives, but rests on the principle of 'balanc[ing] one element against another' (SCAA, 1994a: i) through professional judgement in teacher assessment (SCAA, 1994a) and calibrated aggregation of scores in standard assessment (SCAA, 1993).

As a result of these shifts, National Curriculum assessment has moved from a strong operationalization of the common sense view in the original assessment model, based on denotative measurement and hierarchical progression, to a much weaker operationalization in the revised model, based on connotative measurement and cumulative progression. This article will highlight some central weaknesses of the common sense view of assessment through a discussion of these operationalizations. The argument will be pursued with particular reference to mathematics. Not only is this a subject which has been in the vanguard of the implementation of the National Curriculum; it is also generally regarded as an area in which assessment is relatively unproblematic. For illustrative purposes, evidence will, where possible, be drawn from the development of National Curriculum assessment; in particular, from the internal and external evaluations of the national pilot for assessment in secondary mathematics, which have recently become publicly available (Close *et al.*, 1992; Ruddock *et al.*, 1993). However, to illuminate the issues raised more fully, and to show that they are not peculiar to these circumstances, it will

also be necessary to refer to authoritative prior studies from the UK and elsewhere.

OBJECTIVES AND MEASUREMENT

In mathematics, the first formulation of curricular objectives (DES, 1989) was quickly revised to 'simplify the structure so as to make assessment arrangements more manageable', 'largely by combining existing statements into broader statements of more general application' (DES, 1991a: ii, iii). Küchemann (1990) has commented on the elasticity of the statements of attainment in the original version, stretched still further in the simplified form. To take just one example, at level 6 of the attainment target on number, at least four statements appear to have been compressed into one. 'Pupils should: "understand and use equivalence of fractions and of ratios; relate these to decimals and percentages"; "work out fractional and percentage changes and related calculations"; "calculate using ratios in a variety of situations"; "convert fractions to decimals and percentages and find one number as a percentage of another"' (DES, 1989) became 'Pupils should be able to calculate with fractions, decimals, percentages or ratio as appropriate' (DES, 1991b).

Nevertheless, some clarification of statements of attainment has been provided by the limited number of official examples which accompany each. One example for the 'simplified' statement above, also used to illustrate the corresponding precursor in the original version, is: 'Adapt a recipe for six people to one for eight people.' The evidence in Table 1 is drawn from a study by the Assessment of Performance Unit (APU) (Foxman *et al.*, 1985: 147–8) involving large representative samples of pupils in the last years of primary and compulsory secondary education. The table illustrates the extraordinary variation in success rates over apparently similar items closely matching this example. Similarly, Table 2 shows results from a study carried out by Hart as part of the Concepts in Secondary Mathematics and Science (CSMS) and Strategies and Errors in Secondary Mathematics (SESM) projects (Hart, 1981; 1984: 14) covering a large sample of secondary pupils. These variations in facility indicate that items apparently matching the same example cannot be taken as providing comparable indices of achievement.

For the expert mathematician, of course, the similarity of these items lies in their perceived structure. Within the norms of scholarly mathematics, all are exemplars of the same kind of multiplicative relationship, amenable to solution by correspondingly general methods, such as calculating a unitary ratio or equating two ratio expressions. Indeed, in the APU study, each item is accompanied by a statement of the 'ratio involved' such as '8/12 = x/1500'. This approach to classifying tasks and demarcating objectives draws heavily on the accepted formalization of the discipline. But, as Lakatos argues, this formal

Table 1. Pupil performance on recipe scaling items (APU)

Item	*Pupil success rate*	
	Age 11	*Age 15*
To cook a meal for 12 people I need:		
12 chops		
6 tomatoes		
18 potatoes		
1500 g peas		
I want to cook a meal for 8 people.		
Fill in the amounts of food I need.		
_____ chops	87%	95%
_____ tomatoes	74%	86%
_____ potatoes	34%	61%
_____ peas	21%	46%

Table 2. Pupil performance on recipe scaling items (CSMS)

Item	*Pupil success rate*	
	Age 13	*Age 15*
Onion soup recipe for 8 persons.		
8 onions		
2 pints water		
4 chicken soup cubes		
2 dessertspoons butter		
½ pint cream		
I am cooking onion soup for 6 people.		
How much water do I need?	85%	88%
How many chicken soup cubes do I need?	75%	79%
How much cream do I need?	24%	31%

codification masks the social and individual processes through which mathematical ideas are constructed: 'The phylogenesis and the ontogenesis of mathematical thought cannot be developed without the criticism and ultimate rejection of formalism' (Lakatos, 1976: 4).

This is confirmed by the contrast which emerges between the precise epistemic classification of the recipe problems and the diverse cognitive strategies which they evoke from pupils. Not only did Hart (1981, 1984) find little use for canonical methods in her study of 'ratio' tasks; many pupils could not even be

said to be consistently following an informal method. Rather, most were 'bricoleurs' (Papert, 1980), who 'changed the method they used continuously, adapting to what they saw as the demands of the question. Generally they avoided multiplying by a fraction and tended to build up an answer in small segments, adding them together at the end' (Hart, 1981: 89). In particular, there was much use of strategies combining adding with doubling and halving, often themselves treated as additive processes. Hence items which lent themselves to such additive strategies had considerably higher facilities than those which did not.

One central problem of conceptualization within the common sense view of assessment is now clear. Curricular objectives specified in terms of a formal epistemic analysis of mathematics take no account of the cognitive strategies which pupils actually use to tackle mathematical tasks; of the factors influencing these strategies; and of the ways in which these strategies mature and change. In other words, formal specifications of curricular objectives provide poor criteria for modelling mathematical thinking. Indeed, the uncritical conflation of epistemic and cognitive constructs is a fundamental conceptual weakness within the common sense view; in particular, the way in which epistemic characterizations of a performance are taken to indicate corresponding cognitive capabilities on the part of the performer (Ryle, 1949; Davis, 1990). Indeed, Ryle is all too prescient: 'Presumably, if epistemologists had paid ... attention to arithmetical and algebraical reckonings ... they would ... have used ... analogous arguments to prove the occurrence ... of mental processes.... We might even have been credited with one Faculty of Long Division and another of Quadratic Equations' (1949: 277–8).

Confronted with such cognitively diffuse statements of attainment, the developers of standard assessment have been obliged to refine them covertly, primarily by interpreting each National Curriculum level in terms of a relatively narrow facility band, thus ignoring items which match the rubric of a statement but have too high or too low a success rate for the level at which the statement is placed (Wiliam, 1993: 342). In the differentiated tests of the 1992 pilot, the items operationalizing the illustrative statement were included in the papers set at the two central bands, with an estimated 47 per cent of the cohort presented at Band 3–6 (covering the corresponding National Curriculum levels), and 40 per cent at Band 5–8 (Ruddock *et al.*, 1993: 115). Because the presentation of the items is lengthy and they are structured to require compound responses, their content is summarized in Table 3 rather than reproduced in full: the complete versions can be found in the original test publication (SEAC, 1992). Two sets of success rates are quoted: those from the official external evaluation conducted by the National Foundation for Educational Research (NFER) and Brunel University (Ruddock *et al.*, 1993: Appendix, 7, 10) are from a large and nationally representative sample (p. 115); those from the internal report by the Consortium for Assessment and Testing in Schools (CATS) (Close *et al.*, 1992: Appendix, D19, D21) are from a still larger, but not nationally representative sample (p. 17).

Table 3. Pupil performance on percentage items (SEAC)

Item summaries

A: Two classes are collecting money. Bethan's class want to get £1000. So far they have got 40% of their total. Peter's class want to get £500. What percentage of £500 is the same as 40% of the £1000?
B: Two shops usually sell the same trainers at the same price. There is a sale at both shops. Jane's offers 30% off. Brian's offers ⅓ off. Which shop sells the cheaper trainers? Show your working.

Pupil success rates: Band 3–6

Item	*NFER/B*	*CATS*
Item A	26%	32%
Item B	25%	30%
Both items	10%	15%
At least one item	41%	47%

Pupil success rates: Band 5–8

Item	*NFER/B*	*CATS*
Item A	65%	70%
Item B	68%	75%
Both items	53%	57%
At least one item	80%	88%

At first sight, the item facilities suggest that all is well. Within each band, the proportion of pupils successful on each task is comfortingly similar, encouraging the interpretation that around 25 per cent of lower-band and 65 per cent of upper-band pupils (in the representative sample) have the specified ability. But on examining the pattern of response across the two items, this interpretation disintegrates: within the lower band, the performance of 31 per cent of pupils is inconsistent; in the upper band, the corresponding proportion is 27 per cent. The practical significance emerges in the aggregation of assessments of parallel items. Within the testing procedures, pupils who succeeded on at least one of the parallel items were deemed to have satisfied a statement of attainment. Had success on both items been required, the rate of award of the statement of attainment (in the representative sample) would have fallen dramatically from 41 per cent to 10 per cent in the lower band, and from 80 per cent to 53 per cent in the upper band. Had only one of these items been set (as the evaluation has recommended for future tests, and the political drive to shorten tests will further encourage) the success rate would have been around 25 per cent in the lower band

and 65 per cent in the upper. (In the practice of standard assessment, of course, this is where the idea of the 'correct' facility for the level again comes to the fore. As an anonymous reviewer of this article comments: 'Calibration cannot be conducted independently of the aggregation rule to be used. If we are using a disjunctive ... aggregation model [in which a candidate has to answer just one of a set of items correctly] then we might choose relatively demanding items. However, if a ... conjunctive model is adopted [in which the candidate has to answer all items correctly] then easier items will have to be set. ... Had only one item been set, it would be different from either of the actual ones chosen.')

Similar issues have surfaced in teacher assessment. Reporting on the implementation of the National Curriculum, Her Majesty's Inspectors (HMI, 1991: 26; 1992: 28) record that among the main concerns expressed by teachers about assessment were 'the meaning of "mastery" of a Statement of Attainment [and] how often an ability had to be seen in pupils' work in order to be accepted as evidence of attainment'. In practice, HMI found that most teachers were attempting to record 'at three levels: "work covered", "some degree of understanding" and "mastery"' (HMI, 1991: 25). TGAT, of course, recognized the need to take account of potential variations in performance across assessment tasks employing different formats and set in differing contexts, administered on different occasions and under differing conditions. It was precisely to widen the range of evidence on which assessment was based that TGAT recommended that 'the national assessment system be based on a combination of moderated teachers' ratings and standard assessment tasks' (TGAT, 1988a: para. 63). Widening the range of observation was based on the principle that 'observed responses must be shown to be typical' (TGAT, 1988a: para. 60). However, the implied conception of 'atypical' pupil response is not wholly adequate. Explaining discrepancies in performance in this way may mask an important underlying factor: that the novice mathematician does not perceive problems as structurally similar in the way that the expert does. It is for novices that context and presentation are likely to be at their strongest in influencing strategy; in effect, for such pupils, 'parallel' items may be quite unrelated.

Another example will both illustrate the replicability of the phenomenon, and extend our analysis of it. In France, where the Evaluation du Programme de Mathématiques (EVAPM) has been evaluating a new national mathematics curriculum, Bodin (1989; 1993: 125) has produced very similar evidence. Table 4 shows the success rates of lower secondary pupils on two items clearly matching the National Curriculum statement of attainment. Despite the resemblance of task, presentation and context, 25 per cent of pupils show discrepant responses. Why do such pupils succeed on one item and fail on the other? Omission rates were extremely low, and so corresponding explanations are not plausible. Previous work suggests that some of the discrepancy could indeed be explained in terms of pupils offering a 'careless' response on one occasion but not the other: for example, reporting the reduction in price rather than the new price (Foxman

Table 4. Pupil performance on percentage items (EVAPM)

Item statements	
A: An article costing 400 FF is reduced by 10%. What is the new price of the article after this reduction? B: The price marked on a car is 45,000 FF. The salesman makes a reduction of 5% on this price. What is the new price of the car?	
Pupil success rates	
Item A	41%
Item B	40%
Both items	28%
At least one item	53%

et al., 1985: 140–1). But there are some deeper explanations of why success on Item 4A might be accompanied by failure on Item 4B: calculations with 10 per cent are likely to be familiar and well rehearsed; and the misconception that to find *n* per cent you divide by *n* (which originates as an overgeneralization of the observation that to calculate 10 per cent you divide by 10) produces a 'correct' result in this case. Why, then, might other pupils be successful on Item 4B but not on Item 4A? One plausible explanation is that the simple numeric values in Item 4A influence choice of strategy and means of calculation. First, rather than calculating the reduction of 10 per cent and then subtracting this from the full price, some pupils may calculate the reduced price directly as 90 per cent; and, second, whatever strategy pupils adopt, they may carry out calculations mentally rather than using a written or calculator method. In each case the likelihood of error is increased.

This highlights a further important conceptual weakness of the common sense view. The process of learning cannot be adequately modelled as the once-and-for-all achievement of curricular objectives. Rather, mathematical development is characterized by increasing flexibility and fluency in the deployment of particular strategies; by the capacity to represent these strategies and monitor their use; and by cognitive reorganization which leads both to the rejection of inadequate strategies (such as calculating *n* per cent by dividing by *n*) and to the development of more curtailed strategies (such as calculating the discounted price directly as 90 per cent of the original, rather than calculating the 10 per cent discount and then subtracting it from the original). Performance may be particularly unstable both during the earlier stages of learning, and during periods of cognitive reorganization. It is only as use of a set of interlinked strategies becomes increasingly flexible and fluent that the pupil approaches a

BEST COPY AVAILABLE

degree of mastery likely to produce the stable successful performance on a wide range of structurally related tasks assumed by the common sense view.

The practical difficulties of criterion referencing within the original National Curriculum assessment model are symptomatic of conceptual weaknesses in the common sense view. Far from addressing these weaknesses, the revised assessment model seems set to compound them. The retreat from denotative to connotative measurement is presented as a move 'from the present plethora of detailed statements of attainment to a synoptic description of the key elements that characterize achievement at a particular level' (SCAA, 1994a: i). Assessment no longer seeks to denote the specific achievements of the individual pupil, but to connote such achievements in terms of a compound descriptor which 'best fits the pupil's performance' (SCAA, 1994a: 26). In effect, the new level descriptions sacrifice the epistemic integrity of the individual statements of attainment in favour of an unsubstantiated claim of developmental communality. For example, the statement explored in detail in this section is absorbed into the level description: 'Pupils order and approximate decimals and use these to solve numerical problems and equations of the form $ax^n = 20$, by trial and improvement. Pupils multiply and divide negative numbers in appropriate contexts. Pupils are aware of which number to consider as 100 per cent, or a whole, in problems involving comparisons, and use this to evaluate one number as a fraction or percentage of another. Pupils understand and use the equivalences between fractions, decimals and percentages. They calculate using ratios in a variety of situations. When exploring number patterns, pupils describe in words the rule for generating the nth term of a sequence, where the rule for the nth term is linear. They formulate and solve linear equations with whole number coefficients. They use co-ordinates in all four quadrants to represent mappings, expressed algebraically, interpreting their general features' (SCAA, 1994a: 27).

It is hard to see how this spectacular concatenation addresses the complaint that: 'Many of the criteria set out in the National Curriculum statements of attainment lack precision. It is not surprising that teachers interpret them in different ways and have different views on the knowledge, understanding and skills required at each level' (Dearing, 1993: 40). Indeed, an anonymous reviewer of this paper reports that: 'In launching the new level descriptions [Dearing] noted that it was no longer necessary for a pupil to attain all, or even most, of the separate statements at a level to be judged to have attained that level.' In effect, then, this is not so much a summary descriptor of individual performances as a stereotypical elaboration of the performance of the 'level 6 pupil' in number and algebra. As Dearing says on the statements of attainment that the level descriptions are intended to replace: 'It is clear that their apparent precision is sometimes spurious' (Dearing, 1994a: 58).

PROGRESSION AND SUMMARIZATION

In the original National Curriculum assessment model, the broad assumption was that progression conforms to the hierarchy defined by the levelling of statements of attainment. In principle, a pupil should display a pattern of performance involving consistent success on statements up to the level achieved on each attainment target, possibly partial success on the statements of the following level towards which the pupil is working, and consistent failure on statements beyond that level: 'A pupil assessed as having achieved a given level, say level 2, will have satisfied the criteria for level 2 and will be working towards the criteria for level 3. Progress is marked by achievement of successive levels over time' (TGAT, 1988a: para. 101). In practice, in neither teacher nor standard assessment did the summarization procedures of the national pilot in secondary mathematics conform wholly to this principle. Among the sample of secondary heads of departments questioned in the official evaluation, only 13 per cent reported that they required pupils to show achievement in all statements at an attainment target level to be awarded it, and only 21 per cent that some other proportion of statements was required (Ruddock *et al.*, 1993: 30). In standard assessment, as long as half or more of the statements at an attainment target level had been achieved, 'rollback' of success on statements at higher levels could compensate for the missing statements, with missing statements at lower levels being ignored.

Raw results from the pilot tests can, however, be reanalysed to test the match of pupil performance to the ideal model of hierarchical progression. The consistency measures in Table 5 indicate the extent to which performance accorded with this model (Ruddock *et al.*, 1993: 179). A score of 1 on the consistency index corresponds to pupils scaling perfectly on the hierarchical model; a score of 0 to no consistent scaling whatsoever; with values decreasing steadily from 1 to 0 as the divergence of pupil responses from the model increases (Schagen, 1993). The tabulated figures are sufficiently low to suggest that the

Table 5. Consistency of pupil performance in National Curriculum testing in secondary mathematics (NFER/Brunel)

Domain	*Consistency*	
	Band 3–6	*Band 5–8*
Number	0.53	0.46
Algebra	0.36	0.48
Shape and space	0.54	0.50
Data handling	0.33	0.49

BEST COPY AVAILABLE

substantial discrepancies cannot simply be attributed to deficiencies in the assignment of statements to particular levels in National Curriculum orders and their operationalization in test items, but that they reflect the inadequacy of this view of progression itself.

The evidence of previous studies which have attempted to construct domain hierarchies in mathematics lends support to this conclusion. Construction of workable scales in the CSMS project (Hart, 1981) involved excluding items which did not conform closely to the hierarchical model; even then, anomalous cases continued to arise (O'Reilly, 1990). Perhaps the most meticulous, if small-scale studies that have been carried out are those of Denvir and Brown (1986, 1987). Essentially, their findings suggest that it may be reasonable to model pupil progression over relatively short periods of time in a tightly defined mathematical domain as partially ordered. None the less, some pupils continue to confound even this looser hierarchical structure. And to project the corresponding lattices onto a fully ordered scale involves considerable sacrifice both of informational content and structural coherence.

It is not surprising, then, that the attainment levels awarded to pupils in the national pilot for secondary mathematics prove sensitive to the performance sampling and aggregation procedures adopted (Close *et al.*, 1992: K5). These findings were derived by simulating the effects of alternative procedures using raw results from the pilot tests. As we have seen, in the actual test, two items were set on each statement of attainment, and a pupil was required to be successful on only one of them to be awarded that statement. Table 6 shows the effects of 'shortening' the test, by basing the award of a statement of attainment on performance on the first item only, removing the second opportunity to achieve success on the statement. Overall, around 40 per cent of pupils are awarded a lower subject level as a result of this more restricted sampling of performance (although, of course, in practice, 'easier' items could be set to prevent this effect). Similarly, as we have seen, the actual procedure used to aggregate performance

Table 6. Proportion of pupils attaining a lower test level under sampling based on performance on first statement item only (CATS)

Domain	*Proportion*	
	Band 3–6	*Band 5–8*
Number	39%	53%
Algebra	44%	27%
Shape and space	37%	42%
Data handling	39%	31%
All mathematics	42%	37%

Table 7. Proportion of pupils attaining a lower test level under aggregation requiring success on all statements for level award (CATS)

Domain	*Proportion*	
	Band 3–6	*Band 5–8*
Number	20%	20%
Algebra	1%	4%
Shape and space	12%	35%
Data handling	28%	19%
All mathematics	18%	20%

on statements of attainment into award of an attainment target level allowed 'rollback' of success on statements at higher levels. Table 7 shows the effects of requiring an exact match between statements achieved and level awarded, by removing the opportunity for compensation by 'rollback'. Overall, around 20 per cent of pupils are awarded a lower subject level.

Again we are confronted by the issues identified in the previous section. First, to the extent that learning and development in mathematics involve the progressive consolidation, co-ordination and reorganization of knowledge, with stability of performance across a particular class of tasks achieved only towards the end of this process, then this picture of rather diffuse patterns of performance at the 'leading edge' is not surprising. Second, to the extent that the construction of statements of attainment and their assignment to levels reflect epistemic structures and curricular habits, largely without the benefit of cognitive analyses, the model will tend to be prescriptive rather than descriptive. Finally, and perhaps most importantly, no simple model can provide an interpretable summary of patterns of pupil performance in areas where knowledge is in the process of development and refinement. An attainment scale is likely to provide interpretable information only over those lower levels corresponding to a pupil's expert performance, and those higher levels where the pupil is a complete novice. Attempts to construct a descriptive summary measure in the intervening band of development are simply misconceived.

While Dearing acknowledges that 'as one teacher put it, "learning is a messy business" '; and that 'the assumption that a pupil's development . . . progresses in an orderly way through ten levels is simplistic' in not recognizing 'the progression and regression which marks actual learning' (Dearing, 1993: 26, 40, 41), he seeks to preserve the idea of stepwise progression and the levelled framework for modelling it. 'The ten-level scale is unnecessarily complex and excessively prescriptive. . . . The purposes it was intended to serve are nevertheless sound. Not all of the problems associated with the ten-level scale can be

solved but much can be done to improve it' (Dearing, 1994a: 11–12). In the revised model, then, we find a view of progression as cumulative, based on the principle of 'balanc[ing] one element against another' (SCAA, 1994a: i); assessed not by detailed matching of patterns of achievement to the explicit model of progression, but by holistic judgement in teacher assessment paralleled by the aggregation of scores under psychometric assumptions in standard assessment. Here, the optimistic rider is added that assessment procedures should 'ensure that the award of a test level to a pupil bears a close relation to the pupil's performance on questions targeted at that level' (SCAA, 1993), despite previous experience suggesting that no degree of calibration will make it possible to wring meaning out of levels produced by adding marks awarded across a range of disparate tasks (Wood, 1991: 85).

Dearing's solution, then, effectively inverts the original TGAT model. The primary focus of TGAT is on assessing pupils' specific achievements in terms of individual objectives, with the corresponding scale level intended to provide an overall summary of this pattern of achievement, under the assumption of hierarchical progression. In the Dearing model, the major concern is with assessing pupils' overall achievement in terms of a scale level, with the associated objectives intended to provide an exemplification of a corresponding pattern of specific achievements, under the assumption of cumulative progression.

Here, Dearing is responding to the political drive to 'shorten' tests and produce 'reliable' results, reflected in the progressive shift which had already taken place in testing procedures (SEAC, 1992, 1993; SCAA, 1994b) from the goal of a descriptive profile across different elements of the subject towards that of a single summary index of attainment. Whereas TGAT: '[did] not generally recommend' 'aggregation across profile components in a subject to give an individual pupil a single level or score for reporting' (TGAT, 1988b: 7), essentially on the grounds that such results do not provide interpretable summaries of attainment, Dearing writes that: 'The advice I have received makes it clear that the tests do not need to be so extensive as to yield a separate assessment of the achievement in every attainment target' (Dearing, 1993: 52). For Dearing, then, 'A summative test needs ... to assess a reasonable sample of the pupil's work if it is to provide a valid indicator of achievement. Too short a test runs the risk of giving unreliable information' (1993: 52).

Here again, evidence arising from the pilot tests can throw light on this issue. Table 8 shows that results of classical reliability studies of the test results, again for the two central bands at which the great majority of pupils were presented (Ruddock *et al.*, 1993: 180). Results are analysed first by separate attainment target, and then for mathematics as a whole. Of course, the psychometric assumptions of classical reliability theory are not well matched to the criterion-referenced approach underpinning the design of the National Curriculum tests (Schagen, 1993: 41; Wiliam, 1993: 346). But this is exactly why the resulting evidence is particularly illuminating. In classical testing, items are chosen (and,

Table 8. Reliability of pupil performance in National Curriculum testing in secondary mathematics (NFER/Brunel)

Domain	*Reliability*	
	Band 3–6	*Band 5–8*
Number	0.80	0.74
Algebra	0.65	0.67
Shape and space	0.73	0.81
Data handling	0.66	0.65
All mathematics	0.90	0.91

more particularly, items rejected) so as to match the statistical models entailed by psychometric assumptions. The consequence is often poor content validity. As the official evaluation suggests, the very different approach adopted in designing the National Curriculum tests ensures a good match between curriculum and test content. The tabulated results provide a test of psychometric assumptions against evidence which has not itself been constructed wholly within those assumptions. The (Cronbach alpha) reliability coefficient is an index of the internal consistency of test results, showing the extent to which they can be explained in terms of common factors running across items. Although using different approaches to estimate reliabilities, both evaluations arrive at broadly similar conclusions (Close *et al.*, 1992: 23, K1–K3; Ruddock *et al.*, 1993: 166–80, Appendix, 39–40): that the reliabilities are lower than those which might be expected from tests conducted under psychometric assumptions; but that, interpreted in such terms, they produce estimates of a standard error of measurement of around 0.5 of a level for the attainment target level awarded to a pupil, and 0.25 of a level for the subject level. At this point we must proceed with caution: Wolf and Silver (1993) show that the predominant internal-consistency methods of judging reliability may produce markedly higher estimates than the test–retest methods for which they are usually taken as an appropriate substitute. Assuming, nevertheless, that the internal-consistency estimate is an appropriate one, and that 'errors' of measurement follow the corresponding normal distribution, implies that the rounded subject level awarded to pupils will be 'true' in around 80 per cent of cases.

This approach, however, raises the problem of interpreting a single index of overall attainment. When Dearing suggests that 'there is no complete solution to the problem of differences between pupils' aptitude for learning other than moving to a grouping system which is based entirely on attainment rather than age' (Dearing, 1993: 41), he implies some degree of homogeneity in attainment; and when he talks of a pupil moving through the levels 'in so far as his/her ability

BEST COPY AVAILABLE

allows' (Dearing 1994a: 60) or of 'the risk of the most and least able pupils studying material above or below their abilities' (Dearing, 1993: 41), he invokes the popular idea that some latent trait regulates the learning process. There are, of course, influential assessment models which make such assumptions explicit, notably the Rasch model which, for Wood (1991: 112) 'could be said [to] provide the necessary logical underpinning for classical analysis and the use of the number correct score, which otherwise appear to be motivated by quite pragmatic considerations'. Rasch makes explicit more widely practised assumptions that individual differences in attainment can be explained in terms of some underlying trait, and that assessment tasks have a uniform difficulty across individuals; the specificity of the model is in taking these two factors alone as determining the likelihood of an individual succeeding on an item. Some of the substantial educational and technical criticisms that have been directed at Rasch (Goldstein, 1979) apply much more widely.

First, a unitary construct of mathematical attainment provides a poor basis for modelling pupil achievement. While the reliability results quoted earlier establish that the items of the pilot standard assessment are internally consistent, this does not imply that they are homogeneous or indeed unidimensional (Wood, 1991: 138). The figures shown in Table 9 help to make this more readily apparent: they show estimates of the average inter-item correlation within each domain, calculated using a standard derivation from the reliability coefficient and the number of contributing items (Willmott and Nuttall, 1975: 29). These correlations are low, and consistent with reported factor analyses of mathematics attainment data (Furneaux and Rees, 1978; McIntyre and Brown, 1978) which indicate that the leading component can be expected to explain only around 15 per cent of variance, with many further components needed to account for any reasonable overall proportion. Furneaux and Rees comment on 'the wide dispersion of the mathematics items within the test-space.... Although all the items were very carefully designed, it is clear that about half of them are making no significant contribution to the assessment of any kind of underlying determinant which might be called "mathematical ability"' (1978: 510). McIntyre and Brown conclude: 'The practical possibility of identifying *any* set of dimensions in terms of which pupils' attainments could be described even in a moderately adequate way seems remote. Not only is the variation in performance on any one attainment test commonly describable only in terms of a considerable number of dimensions, but these dimensions are likely to be interpretable, if at all, only in terms of the specific content of the test items' (1978: 45). McIntyre and Brown interpret the principal factor as one of 'general/verbal ability'. In their discussion of examination studies, Willmott and Nuttall (1975: 59) argue that such a general factor, found across a range of subjects, should be interpreted in terms of 'motivation, examination-taking ability and perseverance ... perhaps to as great an extent as general intelligence'; equally, its strong correlation with a test of 'verbal and quantitative aptitude'

Table 9. Average inter-item correlation in National Curriculum testing in secondary mathematics (NFER/Brunel)

Domain	*Inter-item correlation*	
	Band 3–6	*Band 5–8*
Number	0.14	0.14
Algebra	0.13	0.14
Shape and space	0.14	0.21
Data handling	0.11	0.10
All mathematics	0.12	0.14

(Bloomfield, Dobby and Kendall, 1979: 71) suggests that it may also reflect the degree to which pupils have developed core reasoning and learning skills. Against this background, the idea of unitary mathematical attainment, and the interpretation of an aggregate level in such terms, are quite unjustified.

Second, any model needs to take account of pupils' differing approaches to mathematical tasks and, in particular, of the corresponding influence of school experiences and teaching methods. Studying pupils in two schools, both working within the same National Curriculum framework but under very different teaching approaches to mathematics, Boaler (1993) examined responses to number and fraction tasks in a range of contexts. In one school, the teaching approach was based on open-ended activities integrating process and content, with communication and negotiation encouraged between pupils and teachers; in the other school, classroom activity was based on pupils working individually on topic booklets, with investigations set for homework, reinforcing separation of content and process issues. Although overall levels of attainment within the two schools were similar, patterns of performance were very different. In the first school, there was little variation in pupils' procedure and performance between contexts; in the second, there were marked dissimilarities across contexts, with differences in procedure apparently cued by superficial features of the situation. Here, then, similar levels of overall attainment masked very different patterns of performance.

By deflecting attention from specific achievements, the influence of the revised National Curriculum model must be to encourage the characterization of pupils in terms of their overall level of attainment, with the attendant dangers of encouraging level by level teaching (Küchemann, 1990) and ability stereotyping (Ruthven, 1987). Even under the original assessment model there is evidence of teachers interpreting the levelled National Curriculum framework in unduly unitary terms. The official evaluation of the National Curriculum in Mathematics (Askew *et al.*, 1993) reported concern that 'teachers, for ease of working,

[were] decid[ing] a level in mathematics for pupils rather than a level in each target' (138) and that 'anecdotal evidence suggested that some middle and secondary schools were setting and teaching according to the assessed level' (145). Equally, the external evaluation of the national pilot for secondary mathematics assessment (Ruddock *et al.*, 1993: 31–2) found only around half of heads of mathematics reporting that teacher assessment in their department recorded against individual statements of attainment for all attainment targets; elsewhere, at least some recording was at attainment target level only. And, as Wiliam (1993: 343) points out, teachers tend to interpret their evidence so that, justifiably or not, it conforms to the hierarchy. This suggests that, to at least some extent, teacher assessment already embodies the holistic professional judgements adopted in the revised assessment model. Findings from the internal evaluation throw further light on this: Table 10 shows the pattern of correlations between levels awarded in teacher and standard assessment (Close *et al.*, 1992: K4).

At first sight, it seems that all is well: correlations are high, indicating the broad agreement of the different judgements of attainment across the four domains of number, algebra, shape and space, and data handling, and the two methods of teacher assessment and standard assessment. But, as the internal evaluation points out, in a well-functioning assessment system, correlations between assessments of the same domain by different methods (in italic type in the table) should exceed correlations between assessments of different domains by the same method (in bold type). This is not the case, and the evaluation concludes that teacher and standard assessment are not measuring the same constructs. The evaluation attributes this to the sampling of performance within standard assessment; an alternative explanation of the generally higher correlations within teacher assessment is that they reflect the influence of a general 'level' or 'ability' construct on the judgements of some teachers. Indeed, as Wood (1991: 75) argues from a wider review: 'School-based assessment could be about ability, whereas external examinations are indubitably about achievement.'

RECONSTRUCTING NATIONAL CURRICULUM ASSESSMENT

The preceding discussion has identified important weaknesses in the common sense view informing current policies on National Curriculum assessment. Its representation of learning as the stepwise acquisition of curricular objectives must be seriously qualified. First, curricular objectives framed primarily in terms of the formal constructs of the scholarly discipline provide a poor characterization of the developing conceptualizations which pupils bring to the corresponding mathematical tasks. Second, development of the flexibility and fluency of use which constitutes mastery of a conceptual system is necessarily a protracted and recursive process, involving cognitive reorganization and refinement, including

Table 10. Correlations between attainment levels awarded by teacher assessment (TA) and standard assessment (SA) (CATS)

Domain	*Domain*							
	Ma2TA	*Ma3TA*	*Ma4TA*	*Ma5TA*	*Ma2SA*	*Ma3SA*	*Ma4SA*	*Ma5SA*
Ma2 (Number) TA	.							
Ma3 (Algebra) TA	0.88	.						
Ma4 (Shape and space) TA	0.85	0.83	.					
Ma5 (Data handling) TA	0.84	0.83	0.82	.				
Ma2 (Number) SA	*0.78*	0.76	0.79	0.74	.			
Ma3 (Algebra) SA	0.78	*0.77*	0.80	0.73	0.84	.		
Ma4 (Shape and space) SA	0.72	0.71	*0.79*	0.68	0.80	0.79	.	
Ma5 (Data handling) SA	0.72	0.71	0.74	*0.73*	0.79	0.77	0.69	.

BEST COPY AVAILABLE

the integration of conceptual systems encountered at different stages of learning. The representation of assessment as providing robust and interpretable indices of progress is correspondingly problematic. While a form of robustness can be engineered by aggregation under psychometric assumptions, when conducted across a diverse and poorly defined domain, this is at the cost of interpretability.

The original National Curriculum assessment model was a rigorous operationalization of the common sense view: the practical difficulties of implementing it and the inconsistencies apparent in its results are symptomatic of the central weaknesses of that common sense view. In place of the unreasonable ambition of TGAT, Dearing offers a pragmatic accommodation. Rather than identifying and addressing the inadequacies of the common sense view, the revised model simply suppresses the evidence; or, as Dearing puts it, 'It eliminates spurious refinement and thus renders the assessment process more credible' (Dearing 1994a: 62).

While this might be an academically acceptable point at which to conclude this article, it is not a professionally defensible one. The issue of National Curriculum assessment remains: the real challenge is to reconstruct it so as to go beyond the common sense view; to create an approach to assessment which is plausible and comprehensible not only to teachers, but to parents and pupils; and which can realistically be implemented. Consequently, I sympathize with Dearing's desire to find some way of reframing the present system to make it more coherent and viable. Unfortunately, his proposals move in precisely the wrong direction, further entrenching the fundamental flaws of the common sense view. The move from statements of attainment to level descriptions exacerbates lack of match with the process and progress of student thinking. The emphasis on aggregation across the subject further decreases the interpretability of assessments while increasing their spurious precision. Regrettably, the effect of these reforms is likely to be to lower the quality of information provided by National Curriculum assessment.

Raising the quality of assessment information, within the twin constraints of the hierarchical scale and of standard tests, presents a considerable challenge. There is, however, one potentially fruitful but undoubtedly controversial way forward – controversial because it entails abandoning the central icon of the common sense view: the notion of a specific level of attainment achieved by a pupil. Instead, it conceptualizes a pupil's learning as taking place in a *development band*. As we have seen, learning is underpinned by attainments which already form part of a fluent and flexible cognitive system, producing consistently high, if not perfect performance on related tasks. In the development band itself, concepts are still undergoing reorganization and refinement: consequently performance is less consistent, and task context and presentation may have an important influence. Within the National Curriculum framework, then, the development band for an individual pupil within a particular attainment target might be conceived as lying beyond an *expertise level*, operationalized in terms of that pupil's consistently high performance on tasks at this and the

preceding level, and extending upwards to include a *familiarity level*, operationalized in terms of modest (and probably uneven) performance on tasks at this and the preceding level. Neither boundary should be thought of as particularly precise, especially if established on the basis of short tests.

Such an approach to reconceptualizing attainment does not seek to banish the common sense view by fiat, but to expose it to public scrutiny. To the extent that the common sense view is justifiable, the expertise and familiarity levels of pupils will be close if not coincident. But, to the substantial extent that pupils' results will be seen not to conform to the common sense representation, teachers, pupils and parents will be encouraged to revise their views. In particular, such a system can be expected to emphasize the importance both of breaking new ground in learning, and of securing and strengthening material which is already familiar to some degree, exercising a more appropriate backwash on teaching and learning than the present system.

The successful implementation of such a system, however, requires two parallel initiatives. First, it is essential that the definition of attainment targets, and scaling within them, should be improved. This cannot be achieved immediately or easily: we currently lack detailed knowledge of development in some aspects of mathematics, and even where such knowledge is available it has to be interpreted with caution because of the influence of curriculum itself on patterns of development. Rather, a continuing programme of research is needed, aimed at exploiting the results of National Curriculum assessment to influence its development. The goal of this programme should be to explore the interactions between curriculum, classroom and cognition in order to improve the specification and structuring of curriculum domains and refine the means of assessing them.

The second initiative concerns the relation between teacher and standard assessment. The evidence is that present teacher assessment takes similar forms to standard assessment, but often lacks the same degree of careful preparation and detailed scrutiny (HMI, 1991: 25; Ruddock *et al.*, 1993: 28–9). The increasing emphasis on standard assessment through short tests is likely to reinforce these trends. As the evaluation of the National Curriculum in mathematics signals: 'There is . . . a risk of [teacher assessment] becoming little more than a mark on a "mock" test, rather than an ongoing process providing an alternative form of assessment with the potential of being more valid and reliable than a national test score' (Askew *et al.*, 1993: 227). Here the need is for a programme to develop forms of teacher and standard assessment which can be genuinely *complementary* as envisaged by TGAT (1988a: paras 64–80): teacher assessment based on the breadth and depth of evidence necessary to ensure high quality information, and standard assessment providing a means of moderating to common national standards.

In the spirit of Swift, this modest proposal accepts the major constraints imposed by government policy, and represents a strengthening and extension of

recommendations already made by Brown (1991) in an earlier appraisal of National Curriculum assessment. It must be said, however politically unacceptable at present, that short tests are unlikely to provide the most trustworthy or informative means of making assessments. Nor are they likely to have the same influence on the general quality of educational assessment as approaches which provide classroom teachers with materials of high quality for more flexible and regular classroom use, with appropriate moderation. Indeed, it is in such direction that other educational systems, notably that of the United States, are moving, in response to their experience of what have proved to be the debilitating effects of standard testing on mathematical attainment (Lesh and Lamon, 1992; Dossey and Swafford, 1993). The gentle revision of the common sense view through a reconstruction of National Curriculum assessment may hasten the day when this comes to be recognized in the United Kingdom too.

REFERENCES

Askew, M., Brown, M., Johnson, D., Millett, A., Prestage, S. and Walsh, A. (1993) *Evaluation of the Implementation of National Curriculum Mathematics at Key Stages 1, 2 and 3*. London: SCAA.

Bloomfield, B., Dobby, J. L. and Kendall, L. (1979) *Ability and Examinations at 16+*. London: Macmillan.

Boaler, J. (1993) 'Encouraging the transfer of "school" mathematics to the "real world" through the integration of process and content, context and culture'. *Educational Studies in Mathematics* 25(4): 341–73.

Bodin, A. (1989) *Evaluation du Programme de Mathématiques: Fin de Sixième*. Paris: Association des Professeurs de Mathématiques et de l'Enseignement Public.

Bodin, A. (1993) 'What does to assess mean: the case of assessing mathematical knowledge'. In Niss, M. (ed.) *Investigations into Assessment in Mathematics Education*. Dordrecht: Kluwer.

Brown, M. (1991) 'Problematic issues in national assessment'. *Cambridge Journal of Education* 21(2): 215–29.

Close, G., Boaler, J., Hrekow, M., Knight, B. and Symons, K. (1992) *CATS Mathematics Key Stage 3 Report: 1992 Pilot Tests*. London: King's College.

Davis, A. (1990) 'Logical defects of the TGAT Report'. *British Journal of Educational Studies* 38(3): 237–50.

Dearing, R. (1993) *The National Curriculum and its Assessment: An Interim Report*. London: School Examinations and Assessment Council.

Dearing, R. (1994a) *The National Curriculum and its Assessment: Final Report*. London: School Curriculum and Assessment Authority.

Dearing, R. (1994b) Interviewed on Independent Television News, *News at Ten*, 2 February.

Denvir, B. and Brown, M. (1986) 'Understanding of number concepts in low attaining 7–9-year olds'. *Educational Studies in Mathematics* 17: 15–36, 143–64.

Denvir, B., Brown, M. and Eve, P. (1987) *Attainment Targets and Assessment in the*

Primary Phase: Report of the Mathematics Feasibility Study. London: King's College.

Department for Education (1994) *Final Report on the National Curriculum and its Assessment: The Government's Response*. London: Department for Education.

Department of Education and Science (1989) *Mathematics in the National Curriculum*. London: HMSO.

Department of Education and Science (1991a) *Mathematics for Ages 5 to 16 (1991)*. London: HMSO.

Department of Education and Science (1991b) *Mathematics in the National Curriculum (1991)*. London: HMSO.

Dossey, J. and Swafford, J. O. (1993) 'Issues in mathematics assessment in the United States'. In Niss, M. (ed.) *Cases of Assessment in Mathematics Education*. Dordrecht: Kluwer.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P. and Sexton, B. (1985) *Mathematical Development: A Review of Monitoring in Mathematics 1978 to 1982*. Slough: National Foundation for Educational Research.

Furneaux, W. D. and Rees, R. (1978) 'The structure of mathematical ability'. *British Journal of Psychology* 69: 507–12.

Goldstein, H. (1979) 'Consequences of using the Rasch model for educational assessment'. *British Educational Research Journal* 5(2): 211–20.

Hart, K. (1981) 'Ratio and proportion'. In Hart, K. (ed.) *Children's Understanding of Mathematics: 11–16*. London: John Murray.

Hart, K. (1984) *Ratio: Children's Strategies and Errors*. Windsor: NFER/Nelson.

Her Majesty's Inspectorate (1991) *Mathematics Key Stages 1 and 3: A Report by H.M. Inspectorate on the First Year, 1989–90*. London: HMSO.

Her Majesty's Inspectorate (1992) *Mathematics Key Stages 1, 2 and 3: A Report by H.M. Inspectorate on the Second Year, 1990–91*. London: HMSO.

Küchemann, D. (1990) 'Ratio in the National Curriculum'. In Dowling, P. and Noss, R. (eds) *Mathematics versus the National Curriculum*. Basingstoke: Falmer.

Lakatos, I. (1976) *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge: Cambridge University Press.

Lesh, R. and Lamon, S. (eds) (1992) *Assessment of Authentic Performance in School Mathematics*. Washington DC: American Association for the Advancement of Science Press.

McIntyre, D. and Brown, S. (1978) 'The conceptualization of attainment'. *British Educational Research Journal* 4(2): 41–50.

Major, J. (1994) Recorded on Independent Television News, *News at Ten*, 6 January.

O'Reilly, D. (1990) 'Hierarchies in mathematics: a critique of the CSMS study'. In Dowling, P. and Noss, R. (eds) *Mathematics versus the National Curriculum*. Basingstoke: Falmer.

Papert, S. (1980) *Mindstorms*. Brighton: Harvester.

Ruddock, G., Tomlins, B., Mason, K., Holding, B., Reiss, M., Keys, W., Foxman, D. and Schagen, I. (1993) *Evaluation of National Curriculum Assessment in Mathematics and Science at Key Stage 3: The 1992 National Pilot*. London: School Examinations and Assessment Council.

Ruthven, K. (1987) 'Ability stereotyping in mathematics'. *Educational Studies in Mathematics* 18: 243–53.

BEST COPY AVAILABLE

Ryle, G. (1949) *The Concept of Mind*. Harmondsworth: Penguin (paperback edn, 1963).

Schagen, I. (1993) 'Problems in measuring the reliability of National Curriculum assessment in England and Wales'. *Educational Studies* 19(1): 41–54.

School Curriculum and Assessment Authority (1993) *Specification for the Writing of Key Stage 3 Mathematics Tests for Use in 1995–1997*. London: School Curriculum and Assessment Authority.

School Curriculum and Assessment Authority (1994a) *Mathematics in the National Curriculum: Draft Proposals*. London: SCAA.

School Curriculum and Assessment Authority (1994b) *Key Stage 3 Mathematics Tests*. London: SCAA.

School Examinations and Assessment Council (1992) *Key Stage 3 Mathematics Tests*. London: SEAC.

School Examinations and Assessment Council (1993) *Key Stage 3 Mathematics Tests*. London: SEAC.

Task Group on Assessment and Testing (1988a) *A Report*. London: DES.

Task Group on Assessment and Testing (1988b) *Three Supplementary Reports*. London: DES.

Wiliam, D. (1993) 'Validity, dependability and reliability in National Curriculum assessment'. *The Curriculum Journal* 4(3): 335–50.

Willmott, A. S. and Nuttall, D. (1975) *The Reliability of Examinations at 16+*. London: Macmillan.

Wolf, A. and Silver, R. (1993) 'The reliability of test candidates and the implications for one-shot testing'. *Educational Review* 45(3): 263–78.

Wood, R. (1991) *Assessment and Testing*. Cambridge: Cambridge University Press.

Address for correspondence

Dr Kenneth Ruthven, University of Cambridge Department of Education, 17 Trumpington Street, Cambridge CB2 1QA.

THE CURRICULUM JOURNAL
VOL 6 NO 1 SPRING 1995

EDITORIAL

KENNETH RUTHVEN
Beyond common sense: reconceptualizing National Curriculum assessment

ROSEMARY WEBB and GRAHAM VULLIAMY
The changing role of the primary school curriculum co-ordinator

MIKE POND and ALAN CHILDS
Do children learn history from 'Living History' projects?

PAUL GOALEN
Twenty years of history through drama

DI BENTLEY and SAMANTHA DROBINSKI
Girls, learning and science in the framework of the National Curriculum

ROBERT A. SPARKES
No problem here! The supply of physics teachers in Scotland

MARY ELLEN WALSH
Rural students' transitions to secondary school: culture, curriculum and context

REVIEWS

CURRICULUM DOCUMENT UPDATE

Published by Routledge Journals
11 New Fetter Lane London EC4P 4EE

BEST COPY AVAILABLE

THE CURRICULUM JOURNAL VOL 6 NO 1 SPRING 1995 ROUTLEDGE

The Curriculum Journal ISSN 0958–5176

All enquires concerning the submission of articles and correspondence should be addressed to: Mary James, The Editor, University of Cambridge Institute of Education, Shaftesbury Road, Cambridge CB2 2BX, UK.

Tel: 44 (0)1223 69631 Fax: 44 (0)1223 324421

Books for Review should be addressed to: Barry Stierer, Reviews Editor, The Open University, St James' House, 150 London Road, East Grinstead, West Sussex RH19 1ES.

The Curriculum Journal is a peer reviewed journal published three times a year for the Curriculum Association by Routledge Journals, 11 New Fetter Lane, London, EC4P 4EE.

Tel: 44 (0)171 583 9855

Enquiries concerning advertisements should be addressed to Journals Advertising, Routledge at the above address.

© 1995 The Curriculum Association

Except as otherwise permitted under the Copyright, Designs and Patents Act, 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the Publishers or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK. US copyright law is applicable in the USA.

All enquiries concerning subscriptions should be addressed to the Routledge Subscriptions Department, North Way, Andover, Hants SP10 5BE, UK

Tel: 44 (0)1264 342817 Fax: 44 (0)1264 342807

Annual Subscription rates:

UK Institutions £60	**USA Institutions $100**	**ROW Institutions £70**
UK Individual £34	**USA Individual $60**	**ROW Individual £38**

Single issues and back copies £20/$35

Subscription rates include mailing by accelerated surface post.

Special rates are available to members of the Curriculum Association. If you would like to join the Curriculum Association please write to Bob Moon, School of Education, The Open University, Walton Hall, Milton Keynes MK7 6AA. Mark your envelope The Curriculum Journal/The Curriculum Association.

Typeset by Type Study, Scarborough, North Yorkshire
Printed in Great Britain by Redwood Books Limited, Trowbridge, Wiltshire